DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico - Grupo 3

Temas de Procesamiento del Lenguaje Natural - NLP Aplicado

2do Bimestre - 2024

Integrante	LU	Correo electrónico
Antonella Berzzotti	62/20	antonellaberzzotti@hotmail.com
Victoria Klimkowski	1390/21	02vicky02@gmail.com
Federico Hernán Suaiter	37/19	federicoh.suaiter@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (54 11) 4576-3359
http://www.fcen.uba.ar

Abstract

La anotación de datos es una tarea muy importante para el área de NLP (*procesamiento del lenguaje natural*). Para no dejarlo tan sujeto a la subjetividad humana, hay criterios de anotación que dan una guía y lineamientos a cumplir sobre cómo se etiqueta. En este trabajo vamos a medir nuestra performance de anotación y la de ChatGPT 3.5.

Introducción

El discurso de odio en Internet resulta perjudicial debido a que, entre tantas cosas, crea un ambiente hostil y divide a la sociedad. En este trabajo práctico analizaremos y compararemos que ocurre al no tener, y al tener un criterio de anotación bien definido. A su vez evaluaremos la performance entre anotadores y al utilizar una LLM (*Large Language Model*) a través de instrucciones para que realice esta tarea de NLP.

1. Trabajos relevantes del área - En Argentina

Discursos de Odio en Argentina.[1] Un trabajo realizado por el *LEDA* y el *GECID-IIGG* (con apoyo del *CONICET*) que habla sobre el *Índice discurso de odio (DDO)*. En este se busca analizar los discursos de odio según región, cohorte generacional, estudios alcanzados, entre otros. También busca indagar sobre los mismos con respecto a ciertos temas controversiales, tales como el aborto y las vacunas.

2. Data statement

Executive Summary

Este conjunto de datos consta de tweets sobre respuestas a notas de los medios periodísticos más reconocidos y que mayor tráfico generan en Argentina anotados con clasificación sobre la presencia de discurso de odio así como la categoría a la que pertenecen: misóginos, homofóbicos, racistas, clasistas, capacitistas, de odio político, por apariencia o por antecedentes criminales. El idioma principal es el español rioplatense. Utilizamos tres datasets en total. En particular los primeros dos (presentados sin etiquetar), llamados *dataset 1* y *dataset 2*, constan de 100 tweets. El tercero, llamado *dataset 3*, cuenta con aproximadamente 300 tweets; donde solo fueron extraídos 100 tweets. Este último dataset no fue etiquetado por nosotros para medir performance sino que fue el utilizado para definir el prompt que le dimos a la LLM para la anotación de los datasets anteriores.

Curation rational

El discurso de odio en Internet es algo prevalente en plataformas donde la comunicación es anónima (por ejemplo, redes sociales). Esto resulta perjudicial ya que, entre tantas cosas, crea un ambiente hostil y divide a la sociedad.

Los datasets fueron creados con el objetivo de poder detectar la presencia o no de discurso discriminatorio, y en caso de tenerlo, con que características. Este conjunto de datos consta de tweets sobre notas de los medios periodísticos más reconocidos y que mayor tráfico generan en Argentina. En particular el dataset consta de diferentes tweets sobre notas periodísticas (OT) junto con tweets de respuesta a dicha nota (RP).

Annotator Demographic

En la Tabla 1 se presenta información sobre nosotros. Las categorías marcadas con un asterisco (*) indican que son auto-percibidas (o dicho de otra manera, aquellas con las que el anotador se identifica).

Description

En la Tabla 2 figura el número de tweets asociados a cada categoría según nuestras anotaciones para el *dataset 1* y el *dataset 2*. Cabe destacar que el *dataset 1* fue anotado por nosotros previo a leer los criterios de anotación, mientras que el *dataset 2* fue a posteriori. Por su lado, el *dataset 3* fue utilizado para la generación de prompt para que ChatGPT anote los tweets de los primeros dos datasets.

[1] https://www.unsam.edu.ar/leda/docs/Informe-LEDA-1-Discursos-de-odio-en-Argentina-b.pdf

3. Decisiones tomadas

Los datos fueron seleccionados de los *dataset 1 y 2* de manera ordenada asumiendo aleatoriedad ya presente en el corpus. Los últimos 10 de cada uno de los dataset fueron distribuidos a todos los anotadores (de manera tal de poder llevar a cabo una comparación). En el *dataset 3* (para obtener un prompt ideal de ChatGPT) se buscó priorizar una igual cantidad de datos con odio (HATEFUL) y sin odio (UNHATEFUL), así como una distribución balanceada de las diferentes categorías teniendo en cuenta la proporción original. Además incluimos Tweets con más de un tipo de discurso de odio.

Para los tweets anotados por nosotros 3, a veces hubo discrepancias. Esto lo resolvimos mediante voto mayoritario.

Por su parte, el IAA (*acuerdo entre anotadores*) lo calculamos con las calculadoras de alpha de Krippendorf mencionadas en el apéndice. Para esto tomamos, de una columna dada (por ejemplo, "Hateful del primer dataset", o "LGBTI del segundo dataset") los valores nuestros finales y los valores de los anotadores (y análogamente para cada comparación realizada), lo pasamos a un CSV y lo subimos a las calculadoras (ambas para asegurarnos). Tuvimos 108 CSV.

A veces ocurría que la calculadora nos devolvía un alpha *undefined* cuando todos los valores eran iguales. Tomamos la decisión de ponerla en 1 porque no había ninguna discrepancia.

4. Utilizando ChatGPT para NLP

Para realizar anotaciones automáticas decidimos usar `ChatGPT 3.5`. Tal como OpenAI comenta en su *artículo de preguntas frecuentes* [2], esta LLM fue entrenada con una amplia variedad de texto tomado de Internet, que incluye libros, artículos, páginas web y muchas otras fuentes de información disponibles públicamente. En base a esto, decidimos comparar el enfoque de **Zero-Shot** (sin ejemplos) y el enfoque **Few-Shot** (con ejemplos). Todo esto se realiza especificando el QUE debe de hacer y como clasificar.

En la construcción del prompt (instrucción) de **Zero-Shot** primero pasamos por dos versiones previas. La primera consistió en utilizarlo de manera general. En este caso le pasábamos a ChatGPT el título de la nota, el RP y le pedíamos que en base a eso nos dijera si es un discurso de odio y a que población estaba destinado. El problema: en ocasiones asignaba categorías de discurso de odio distintas a las que nosotros utilizamos.

En consecuencia decidimos ir por un enfoque intermedio y luego por uno final con criterios detallados. Este último fue el utilizado en el prompt de **Zero-Shot**. Sumado a lo anterior se le pidió que (en caso haber discurso de odio) clasifique según nuestras categorías y entre paréntesis se le brindó una pequeña descripción de qué es cada una de ellas. A su vez se le añadió el siguiente criterio: 'en caso de que no quede claro que haya un mensaje discriminatorio (o parezca de carácter difuso) etiquetar como no discriminatorio'.

En la construcción del prompt de **Few-Shot** se diseño una única versión en la que a ChatGPT se le brinda el mismo prompt que **Zero-Shot** pero se le agregan diferentes ejemplos extraídos del criterio de anotación con los que contábamos nosotros para realizar el etiquetado del segundo dataset sobre las diferentes categorías.

5. Comparación de agreement

Para calcular el IAA utilizamos el **coeficiente alpha de Krippendorff**, el cual es una medida estadística utilizada para medir el grado de acuerdo. Utilizamos la misma para comparar entre las siguientes anotaciones.

5.1. Entre nuestras anotaciones

En la Tabla 3 podemos ver que el agreement entre nosotros tres en general es más alto al contar con criterios de anotación proporcionados. Aunque los resultados empeoran particularmente en las categorías Women y LGBTI (probablemente por la naturaleza diferente de los datasets y la influencia de otras subjetividades), en promedio se mejora el alpha de las distintas categorías así como el de la categoría Hateful.

[2] `https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed`

5.2. Nuestras anotaciones vs Anotadores

En la Tabla 4 observamos nuevamente que nuestras anotaciones son más similares a las de los anotadores cuando los criterios de anotación están claramente definidos. Sin embargo, otra vez, esto no se aplica a todas las categorías pero en promedio se mejora el alpha. También se mejora el alpha de Hateful.

5.3. Anotadores vs GPT

En la Tabla 5, al comparar los resultados para el primer dataset con **ZS** (**Zero-Shot**) y **FS** (**Few-Shot**) podemos ver que con **FS** obtuvimos una mejora en todas las categorías.

En cambio, al observar los resultados del segundo dataset, en algunas categorías se obtuvo un peor resultado. Por ejemplo: en Class pasamos de tener 0.791 a 0.653. También algo a notar es que en la categoría Disabled si bien en ambos casos es negativo, cuando trabajamos con **FS** se puede ver una pequeña mejora.

En ambos datasets, **FS** resultó tener un promedio mayor de alphas entre categorías y en Hateful que **ZS**.

5.4. Nuestras anotaciones vs GPT

En la Tabla 6 lo que más destaca es el bajo valor obtenido en Hateful con el primer dataset de **FS**. En particular tuvimos muchas discrepancias contra ChatGPT. Consideramos que esto se debe a que nuestras anotaciones también difieren con la de los anotadores en el primer dataset. Esto se ve enmendado luego de leer el criterio de anotación, mostrando como el mismo permitió obtener anotaciones más similares a las que ChatGPT (y los anotadores) realizaron.

Al igual que antes, en ambos casos **FS** resultó tener un promedio mayor de alphas entre categorías que **ZS** (aunque el mismo fuese muy bajo).

6. Conclusiones

Luego de realizar el trabajo, pudimos notar que para realizar anotaciones se trabaja mejor cuando se tiene un criterio dado. Esto se puede observar en la Tabla 3 y en el Tabla 4 (además de lo comentado al comparar nuestras anotaciones contra ChatGPT).

También nos llamó la atención que en ciertas categorías, el valor alpha de Krippendorff resultó negativo, indicando discrepancias sistemáticas que superan lo esperado por probabilidad (es decir, que las discrepancias observadas son mayores o más frecuentes de lo que se esperaría simplemente por azar).

A su vez en las comparaciones contra GPT, pudimos ver que en comparación con los anotadores tiene bastante discrepancia en el segundo dataset, mientras que en el primero presenta unos valores relativamente altos (tanto en **ZS** como **FS**). En cambio, contra nuestras anotaciones se puede ver como la discrepancia se aplica en ambos datasets.

Creemos que ChatGPT tiene la tendencia de tergiversar o sobre-analizar la información que se le brindó. Por ejemplo, en un momento una noticia que habla sobre como un preso cae de una altura elevada en un escape y sufre un accidente, uno de los comentarios hace alusión a su condición en la cárcel y ChatGPT lo cataloga como CRIMINAL y DISCAPACIDAD; esto último por estar burlándose de un, ahora, lisiado.

Entrenar una LLM para realizar un trabajo puntual (en este caso, el detectar discursos de odio en una red social bajo un contexto dado) es un trabajo que requiere mucho análisis y varias iteraciones.

Vimos que mediante un buen prompt que deje en claro los objetivos que se busca lograr y brinde información extra a través de ejemplos sobre cómo cumplir dicho objetivo, se pueden obtener buenos resultados. Consideramos que, tal como describimos en nuestra presentación del paper sobre `InstructGPT`, es necesario realizar varios análisis sobre diferentes casos, donde idealmente produciríamos (luego de diferentes iteraciones, mecanismos de recompensa, algoritmos de entrenamiento, etc.) distintos modelos afinados relacionados a lo que nosotros deseamos.

7. Apéndice

Calculadoras utilizadas para calcular el **coeficiente alpha de Krippendorff**:

- Recal3 - ReCal: reliability calculation for the masses - LINK
- Krippendorff's Alpha Calculator - K-Alpha Calculator - LINK

Hojas de cálculo:

- Resultados GPT3.5 (**prompt** junto a resultados del dataset 3 extraído) - LINK
- Anotaciones del dataset 1 y 2 — por nosotros - LINK
- Anotaciones del dataset 1 y 2 — por ChatGPT (**Zero-Shot y Few-Shot**) - LINK

CSV:

- Carpeta con los CSV utilizados para calcular el **alpha de Krippendorff** - LINK

Tablas y Figuras

	Antonella	**Victoria**	**Federico**
Edad	24	21	24
Género*	Femenino	Femenino	Masculino
Etnia*	Caucásica	Caucásica	Caucásica
Nacionalidad	Argentina	Argentina	Argentina
Orientación sexual*	Heterosexual	Bisexual	Heterosexual
Lenguaje utilizado	Español rioplatense	Español rioplatense	Español rioplatense
Dominio del lenguaje	Nativo	Nativo	Nativo
Entrenamiento particular	Ninguno	Ninguno	Ninguno

Tabla 1: Información sobre nosotros

	Primer dataset	**Segundo dataset**	**Tercer dataset (extraído)**
Women	4	5	4
LGBTI	0	7	10
Racism	5	11	11
Class	2	8	2
Politics	37	15	5
Disabled	2	2	3
Apprearence	3	10	12
Criminal	2	10	10

Tabla 2: Descripción del dataset - cantidad de tweets de odio asociados por característica

	Primer dataset	**Segundo dataset**
Women	1	0.482
LGBTI	1	0.71
Racism	0.284	1
Class	0.284	1
Politics	0.386	1
Disabled	1	1
Apprearence	1	1
Criminal	1	1
Promedio	0.744	0.899
Hateful	0.768	1

Tabla 3: Comparación de alpha de Krippendorff entre nuestras tres anotaciones

	Primer dataset	Segundo dataset
Women	0.741	0.541
LGBTI	1	0.541
Racism	0.904	0.669
Class	0.19	0.427
Politics	0.072	0.478
Disabled	0.316	0.658
Apprearence	-0.036	0.717
Criminal	-0.021	0.779
Promedio	0.396	0.601
Hateful	0.163	0.680

Tabla 4: Comparación de alpha de Krippendorff entre nuestras anotaciones y los anotadores

	Primer dataset ZS	Segundo dataset ZS	Primer dataset FS	Segundo dataset FS
Women	0.388	0.031	0.658	0.471
LGBTI	1	0.647	1	0.647
Racism	0.594	0.587	0.694	0.719
Class	0.471	0.791	0.713	0.653
Politics	0.126	0.354	0.459	0.483
Disabled	0.558	-0.021	0.741	-0.0015
Apprearence	0.316	0.418	0.558	0.054
Criminal	0.658	0.324	0.658	0.668
Promedio	0.607	0.391	0.685	0.462
Hateful	0.302	0.519	0.727	0.561

Tabla 5: Comparación de alpha de Krippendorff entre los anotadores y ChatGPT. **ZS** (Zero-Shot). **FS** (Few-Shot)

	Primer dataset ZS	Segundo dataset ZS	Primer dataset FS	Segundo dataset FS
Women	0.388	0.223	0.658	0.223
LGBTI	1	0.647	1	0.647
Racism	0.498	0.695	0.591	0.756
Class	0.264	0.427	0.316	0.294
Politics	0.272	0.408	-0.011	0.461
Disabled	0.388	-0.01	0.316	-0.005
Apprearence	-0.015	0.296	-0.021	0.118
Criminal	0.264	0.427	0.388	0.668
Promedio	0.382	0.389	0.405	0.395
Hateful	0.218	0.443	0.054	0.402

Tabla 6: Comparación de alpha de Krippendorff entre nuestras anotaciones y ChatGPT. **ZS** (Zero-Shot). **FS** (Few-Shot)

Volver a Sección 5

Prompts

Zero-Shot

"Dada esta noticia: $< noticia >$
Y esta respuesta: $< respuesta >$
Solamente decir si la respuesta a dicha noticia es un discurso de odio y de así serlo, decir a cual o cuales de las siguientes categorías pertenece: MUJER (si es por su sexo. Insultos sin referencia particular a su condición de mujer no son validos), LGBTI (si es por género o identidad sexual), RACISMO (Xenofobia, racismo), POBREZA (por situación socioeconómica o barrio de residencia), POLITICA (por su opinión o ideología política. No incluye acusaciones de corrupción ni expresiones que tratan de inútiles a funcionarios), DISCAPACIDAD (por tener discapacidades, problemas salud mental o de adicciones), ASPECTO (por su aspecto o edad) y/o CRIMINAL (por sus antecedentes o situación penal (presos)). Si no queda claro que haya un mensaje discriminatorio o parece de carácter difuso, etiquetar como no discriminatorio"

Few-Shot

"Dada esta noticia: $< noticia >$
Y esta respuesta: $< respuesta >$
Solamente decir si la respuesta a dicha noticia es un discurso de odio y de así serlo, decir a cual o cuales de las siguientes categorías pertenece: MUJER (si es por su sexo. Insultos sin referencia particular a su condición de mujer no son validos), LGBTI (si es por género o identidad sexual), RACISMO (Xenofobia, racismo), POBREZA (por situación socioeconómica o barrio de residencia), POLITICA (por su opinión o ideología política. No incluye acusaciones de corrupción ni expresiones que tratan de inútiles a funcionarios), DISCAPACIDAD (por tener discapacidades, problemas salud mental o de adicciones), ASPECTO (por su aspecto o edad) y/o CRIMINAL (por sus antecedentes o situación penal (presos)). Si no queda claro que haya un mensaje discriminatorio o parece de carácter difuso, etiquetar como no discriminatorio.
Ejemplos de discurso discriminatorio hacia las mujeres por la noticia - 'Nati Jota furiosa por los comentarios que recibe en las redes': (Ejemplo de odio hacia las mujeres) 'Pero si sos de plástico nena! (opina de manera denigratoria de su apariencia)', 'Flor de gato!', 'Miauuu!', '$< emojis >$', 'A esta sólo se la conoce por su cuerpo y ahora se hace la santa. Andá a estudiar', 'Le damos hasta que San Lorenzo vuelva a Boedo', 'Y esta rubia tarada quién es?'.
Ejemplo de odio hacia LGBTI por la noticia - 'Anibal Pachano sobre la cuarentena: este virus nos está destruyendo a los actores'. 'Qué asco este sujeto', 'Y a este trolo quién le pidió su opinión?', 'Me desagrada'.
Ejemplos de odio por racismo ante la noticia - 'Rescatan en China a cuarenta gatitos bebé que iban a ser utilizados en restaurantes': 'Chinos asquerosos', 'Malparidos! Chinos de mierda', 'Sigan desparramando pestes hijos de puta!', 'Por qué no se comen entre ellos?', 'País horrible y enfermo', 'Estos chinos nos diseminan su peste por todo el mundo'.
Ejemplos de odio de clase por la noticia 'Presupuesto: aumentó el gasto en planes asistenciales durante la pandemia': 'Basta de mantener vagos!', 'Cansada de los planeros', 'Che laburar estos atorrantes ni en pedo no?', 'PARASITOS'.
Ejemplos de odio politico ante la nota - 'Aumentó el gasto en planes asistenciales durante la pandemia': 'BASTA ZURDOS DE ROBARNOS', 'Bolcheviques de mierda'.
Ejemplos de odio por aspecto ante la nota Luis Brandoni - 'No convoqué el banderazo': 'Viejo de mierda!', 'Qué decrépito impresentable que es este señor', 'Estás gagá, pelotudo'.
Ejemplos de odio criminal por la nota - 'Muere un delincuente tras un enfrentamiento con la policía': ' Uno menos!', 'Excelente!', '$< emojis >$', 'Que pena, pobrecito'.
Ejemplos de odio de discapacidad por noticias sobre Patricia Bullrich: 'Largá la$< emoji/s >$Pato', 'Borracha hdp'"